9.1: Estimating a Population Proportion

(confidence intervals for a proportion)

Recall: A *parameter* is a numerical summary of a population; a *statistic* is a numerical summary of a sample. (For example, the population mean and population standard deviation are parameters; the sample mean and sample standard deviation are statistics.)

Definition: A *point estimate* is the value of a statistic that estimates the value of a parameter.

Definition: A confidence interval for an unknown parameter is an interval of numbers generated by a point estimate for that parameter.

Definition: The *confidence level* (usually given as a percentage) represents how confident we are that the confidence interval contains the parameter.

If a large number of samples is obtained, and a separate point estimate and confidence interval are generated from each sample, then a 95% confidence level indicates that 95% of all these confidence intervals contain the population parameter.

A confidence interval is obtained by placing a *margin of error* on either side of the point estimate of the parameter.

In other words, the confidence interval consists of: Point estimate $\pm$ margin of error

Point estimates for the population proportion:

The point estimate of the population proportion p is the sample proportion $\hat{p}$.
The point estimate of the mean of the sampling distribution of the sample proportions is $\mu_{\hat{p}} = \hat{p}$.
The point estimate of the standard deviation of the sampling distribution of the sample proportions is

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

confidence interval for mean: $\bar{x} \pm t_{\frac{\alpha}{2}} s_{\bar{x}}$

$= \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

So, for every sample, the sample proportion will be in the center of the confidence interval. If we use E to indicate the margin of error, the confidence interval is $\hat{p} \pm E$, or $\left(\hat{p}-E, \hat{p}+E\right)$

If we use the sample proportion $\hat{p}$ as a starting point, we should be able to write the confidence interval as $\left(\hat{p}-z_c\sigma_{\hat{p}}, \hat{p}+z_c\sigma_{\hat{p}}\right)$, where $\sigma_{\hat{p}}$ is the standard deviation of the sampling distribution of the sample proportions, and z_c is a multiplier that tells us how many standard deviations (of the sampling distribution of the sample proportions) lie between the sample proportion $\hat{p}$ and the edge of the confidence interval. We call this z_c the *critical value* for a *z*-score in the sampling distribution of the sample proportions.

Constructing the confidence interval for the proportion:

Procedure:

$n\hat{p}(1-\hat{p}) \geq 10$

1. Verify that $\cancel{n(1-\hat{p}) \geq 10}$ and that the sample is no more than 5% of the population.

2. Determine the confidence level, $1-\alpha$.

3. Determine the critical value $z_{\alpha/2}$ (using the standard normal table).

4. Use the sample proportion to estimate the standard deviation of the sampling distribution of the sample proportions:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$ (standard error)

5. Multiply the critical value $z_{\alpha/2}$ by the estimated standard deviation $\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ to obtain the margin of error.

6. Add and subtract the margin of error from the sample proportion to obtain the lower and upper bounds of the confidence interval:

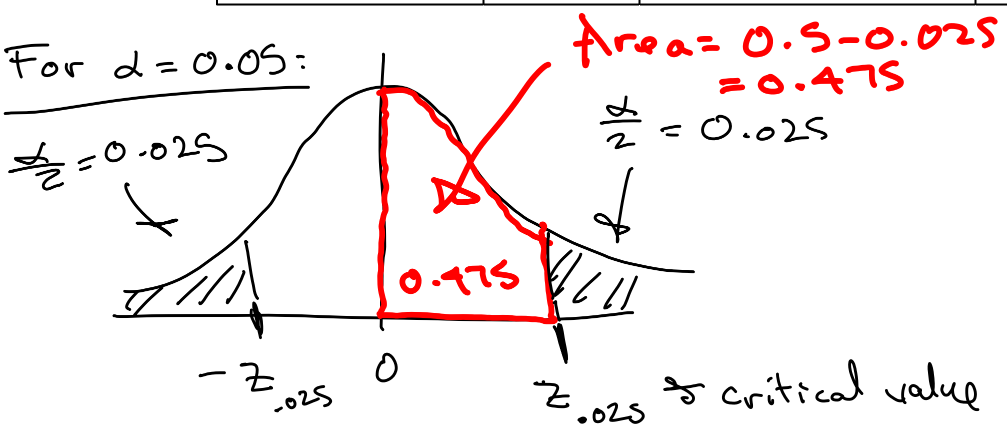
   Lower bound: $\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

   Upper bound: $\hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Useful critical values of *z*:
(You can use these instead of looking in the table every time).

Confidence Level	α	Area in each tail, $\alpha/2$	Critical value $z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.575

For $\alpha = 0.05$:



Look up area $= 0.475$ in z-table.
What z-score does it correspond to?

$z = 1.96$

Example 1: In a random sample of 537 Americans, 173 indicated that they frequently ate peanut butter. Construct and interpret the 90% and the 95% confidence intervals for the proportion of Americans who frequently eat peanut butter.

$n = 537$, Proportion of Americans who eat peanut butter $= p$

Sample proportion: $\hat{p} = \frac{173}{537} \approx 0.322$

Check normality: Is $n\hat{p}(1-\hat{p}) \geq 10$

$537(0.322)(1-0.322)$

$= 537(0.322)(0.678)$

$\approx 117 \geq 10$ yes!

Find standard error:

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.322(0.678)}{537}} \approx 0.020163$$

See next page

Sample size needed to estimate the population proportion within a given margin of error:

When constructing a confidence interval about the sample proportion $\hat{p}$, the margin of error is

$$z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We can solve this for *n*:

In order to calculate the *n* needed, we need an estimate of the population proportion *p*. If such an estimate is available (perhaps from a prior study), we can use the above formula to calculate *n*. If not, we use the very conservative assumption that $\hat{p} = 0.5$, which gives us the maximum possible value for $\hat{p}(1-\hat{p})$, which is $(0.5)(0.5) = 0.25$.

<u>Required sample size for estimation of the population proportion</u>:

For a specified α associated with a confidence level, the sample size required to estimate the population proportion within a margin of error E is

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2,$$

where $\hat{p}$ is an estimate of the population proportion p.

If no estimate for the population proportion is available, we should use a sample size of at least

$$n = 0.25\left(\frac{z_{\alpha/2}}{E}\right)^2.$$

In both cases, because the calculated n is considered a minimum threshold, we round the calculated value of n <u>up</u> to the nearest whole number *above*.

<u>Example 2:</u> A pollster wishes to estimate the percentage of likely voters who support Candidate A. Based on earlier polls, the pollster expects the candidate's level of support to be approximately 38%. What sample size should be obtained if the pollster wishes to estimate the candidate's support level within a margin of error of 3 percentage points, with 95% confidence?

$\hat{p} = 0.38$ $\qquad \alpha = 0.05 \Rightarrow z_{.025} = 1.96$

$\alpha = 0.05$

$E = 0.03$

$$n = \hat{p}(1-\hat{p})\left(\frac{z_{.025}}{E}\right)^2$$

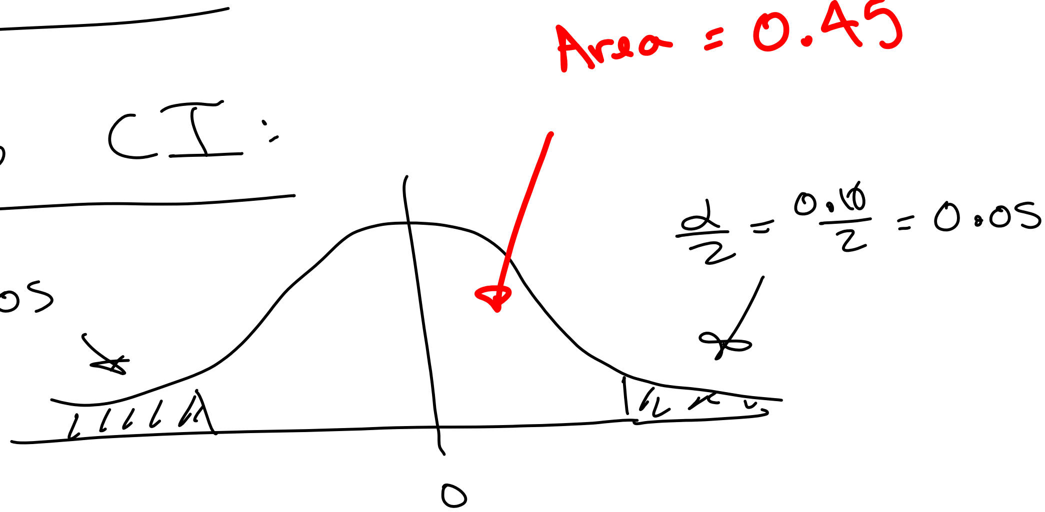
$$= 0.38(0.62)\left(\frac{1.96}{0.03}\right)^2$$

$$\approx 1005.65 \Rightarrow \boxed{\text{need a sample size of at least } 1006}$$

Example 1 cont'd:

For 90% CI:

$\alpha = 0.10$ $\quad \frac{\alpha}{2} = 0.05$



when constructing confidence intervals, we always shade the area corresponding to a 2-tailed test.
(so $\frac{\alpha}{2}$ in each tail)

Look up area = 0.45 in table,
Find the critical value of z that corresponds to this area. $\quad z_{.05} = 1.645$

Upper bound:

$$\hat{p} + z_{.05}\,\sigma_{\hat{p}} = 0.312 + 1.645(0.020163)$$
$$= 0.355$$

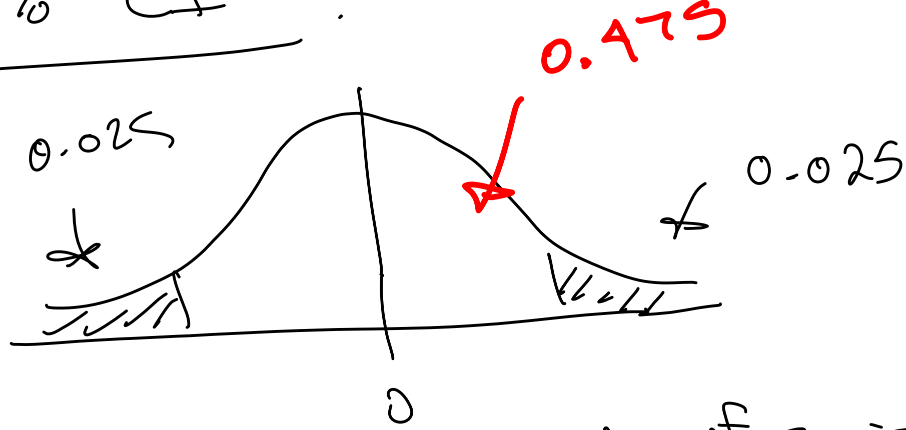
Lower bound:

$$\hat{p} - z_{.05}\,\sigma_{\hat{p}} = 0.322 - 1.645(0.020163)$$
$$= 0.288$$

$$\boxed{90\%\ \text{CI}: (0.288, 0.355)}$$

Find 95% CI :

$\alpha = 0.05$



From z-table, critical value of z is

$$z_{.025} = 1.96$$

Upper bound: $\hat{p} + z_{.025}\sigma_{\hat{p}} = 0.322 + 1.96(0.02016\overline{3})$

$$\approx 0.362$$

Lower bound: $\hat{p} - z_{.025}\sigma_{\hat{p}} = 0.322 - 1.96(0.02016\overline{3})$

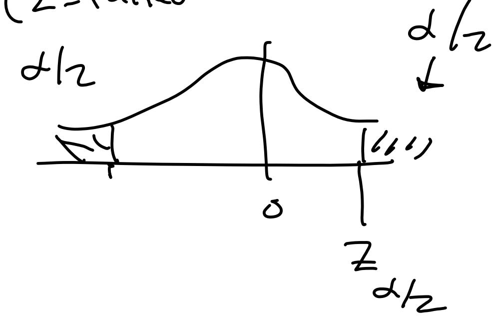
$$= 0.282$$

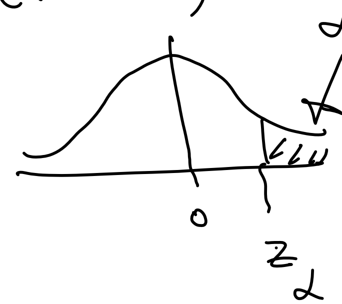
$$\boxed{\text{95\% CI: } (0.282, 0.362)}$$

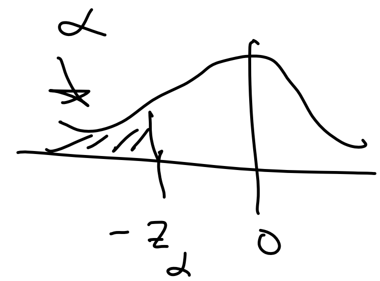
10.2: Hypothesis Tests for a population proportion

For a hypothesis test on a proportion, we have:

$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$
$H_1 \; p \neq p_0$	$H_1: p > p_0$	$H_1 \; p < p_0$
(2-tailed test)	(1-tailed)	(1-tailed)







Use z-table to find the critical value of z.

$z_{\frac{\alpha}{2}}$ or z_α

Compute test statistic:

$$z_0 = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

If z_0 is more extreme than the critical value $z_{\frac{\alpha}{2}}$ or z_α, then you reject H_0.

(If it's in the rejection region, you reject H_0)

10.2 #16 Manufacturer of a drug claims that more than 94% of patients taking the drug are healed.

In clinical trials, 213 of 224 patients taking the drug were healed. Test the manufacturer's claim at the $\alpha = 0.01$ level of significance.

corresponding to confidence level of 99%

$H_0: p = 0.94$

$H_1: p > 0.94$

We'll finish next time!